Mitigating Membership Inference Vulnerability in Iterative Federated Clustering Algorithm

Kangsoo Jung
INRIA, Ecole Polytechnique
Palaiseau, France
gangsoo.zeong@inria.fr

Sayan Biswas
EPFL
Lausanne, Switzerland
sayan.biswas@epfl.ch

Catuscia Palamidessi
Inria, Ecole Polytechnique
Palaiseau, France
catuscia@lix.polytechnique.fr

Abstract

Federated Learning (FL) has emerged as a promising paradigm for collaborative model training without the need to share clients' personal data, thereby preserving privacy. However, the non-IID nature of the clients' data introduces major challenges for FL, highlighting the importance of personalized federated learning (PFL) methods. In PFL, models are typically trained to cater to specific feature distributions present in the population data. A notable method for PFL is the Iterative Federated Clustering Algorithm (IFCA), which mitigates the concerns associated with the non-IID-ness by grouping clients with similar data distributions. While it has been shown that IFCA enhances both accuracy and fairness, its strategy of dividing the population into smaller clusters increases vulnerability to Membership Inference Attacks (MIA), particularly among minorities with limited training samples. In this paper, we introduce IFCA-MIR, an improved version of IFCA that integrates MIA risk assessment into the clustering process. Allowing clients to select clusters based on both model performance and MIA vulnerability, IFCA-MIR achieves an improved performance with respect to accuracy, fairness, and privacy. We demonstrate that IFCA-MIR reduces the risk of MIA by up to 5.6× compared to the original IFCA while maintaining comparable model accuracy and fairness.

Keywords

Personalized federated learning, Membership inference attack, Fairness

1 Introduction

Recent advancements in machine learning technologies have positioned artificial intelligence (AI) as a cornerstone of innovation across various sectors, driving progress in healthcare, finance, education, and beyond. However, concerns about social risks, particularly privacy violations, have become increasingly prominent. These concerns underscore the need for having privacy-preserving AI solutions capable of safeguarding the sensitive data of the clients while maintaining model performance.

Federated Learning (FL) [14] has become a key approach for enabling collaborative model training across multiple clients without the need to share raw data. In FL, participants locally train models on their devices using their own data, sharing only model updates with a central server responsible for aggregation and orchestrating the learning process. This localized and collaborative approach to training machine learning models ensures that raw data remains on clients' devices, effectively minimizing the risk of privacy breaches during the training process.

Despite these advantages as far as privacy is concerned, FL faces significant challenges when applied to real-world scenarios due to the non-IID nature of population data stemming from the clients' diverse behaviors, preferences, and environments. Such feature heterogeneity of data distribution present in the population often impedes model convergence and deteriorates model accuracy. In this context, a one-size-fits-all approach to obtaining a model is inadequate, as it fails to capture the nuanced patterns within each group. Therefore, tailoring models to accommodate these variations is crucial for enhancing predictive accuracy and ensuring fairness across different client groups.

To address the challenge imposed by the non-IID-ness of the training data, Personalized Federated Learning (PFL) approaches have been proposed. These typically cluster participants based on similar data distributions and train models tailored to each group. A cutting-edge method for PFL is the Iterative Federated Clustering Algorithm (IFCA) [9], which employs a clustering-based approach for PFL. In IFCA, the server initializes multiple models and distributes them to participating clients. Each client then selects the best model that minimizes their local loss, performs local training using the selected model, and sends the locally optimized models back to the server. The server aggregates and updates the models, and redistributes the refined models to the clients. This iterative process continues until convergence.

By clustering models according to groups of clients with similar data distributions, IFCA inherently improves both model accuracy and group fairness compared to traditional FL paradigm [6]. Moreover, it has been shown that IFCA can be incorporated with local obfuscation techniques to foster group-level formal privacy guarantees [7].

However, despite these improvements in accuracy, group privacy, and fairness, IFCA has a notable drawback: it becomes increasingly vulnerable to Membership Inference Attacks (MIA) as it partitions the overall dataset into smaller clusters for personalized training. MIA is a significant privacy threat in machine learning, aiming to determine whether a particular data point was used in a model's training set. In IFCA, this vulnerability is particularly pronounced among unprivileged groups or minorities, where smaller training datasets increase the likelihood of successful MIA.

In this paper, we propose an improved version of IFCA called IFCA-MIR (Iterative Federated Clustering Algorithm with Membership Inference Robustness) to address these limitations. Unlike the

ARTMAN '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1909-7/25/10
https://doi.org/10.1145/3733821.3763022

original IFCA, which selects models solely based on empirical loss, IFCA-MIR integrates MIA risk assessment into the clients' model selection process. By balancing both loss and MIA risk, IFCA-MIR reduces the exposure to MIA while maintaining comparable accuracy and fairness. This approach allows privacy-sensitive clients to choose safer clusters, effectively mitigating MIA vulnerability without significantly compromising model accuracy. Our method is designed to achieve a better trade-off between accuracy and privacy, addressing the original IFCA's vulnerability to MIA.

The key contributions of this paper are as follows:

- We demonstrate the MIA vulnerability of the original IFCA algorithm, particularly in minority groups with smaller training datasets. To the best of our knowledge, this is the first study to evaluate the MIA vulnerability of clustering-based PFL algorithms.
- We propose IFCA-MIR, an enhanced IFCA algorithm that integrates MIA risk into the model selection process, balancing empirical loss with privacy considerations.
- We show that IFCA-MIR preserves the theoretical convergence guarantees of IFCA and empirically evaluate the performance of IFCA-MIR in terms of MIA robustness, fairness, and accuracy through extensive experiments on the MNIST, FEMNIST, and CIFAR-10 datasets. Our results demonstrate that IFCA-MIR reduces the number of MIA violations by up to 5.6 times compared to the original IFCA, while maintaining comparable model accuracy and fairness."

The rest of this paper is organized as follows: In sections 2 and 3, we introduce related works and preliminaries. Section 4 details the proposed IFCA-MIR algorithm and analyzes its formal convergence guarantees. Section 5 presents experimental results demonstrating that the proposed algorithm reduces MIA vulnerability without significantly compromising accuracy. Section 6 discusses key insights that require further exploration. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Related works

2.1 Federated learning

Privacy. FL [14] is a framework that allows multiple clients to collaboratively train models without sharing raw data. However, despite its inherent privacy-preserving design due to the localization of the clients' data, FL has been shown to be still vulnerable to various privacy-invasive attacks, as adversaries can exploit shared model updates to extract sensitive information. Studies have demonstrated that malicious clients, servers, or external attackers can utilize gradient inversion techniques [8], differential data leakage attacks [25], and membership inference attacks [12, 24] to reconstruct private training data or infer membership. These vulnerabilities reveal the limitations of standard FL protocols in safeguarding client privacy, emphasizing the need for enhanced privacy-preserving mechanisms.

Personalization. In addition to privacy concerns, one of the critical challenges in FL is the presence of non-IID data across clients. Several lines of work [9, 13, 18, 21] have begun exploring PFL approaches, which aim to tailor models to individual clients' unique data distributions. However, these methods primarily focus on improving model accuracy and convergence without adequately addressing the accompanying privacy risks. In particular, the non-IID nature of data amplifies the risk of privacy attacks, as attackers can exploit unique aspects of the clients' data distributions to infer sensitive information about them. Therefore, there is a growing need for PFL algorithms that not only address non-IID data issues but also incorporate privacy-preserving measures.

2.2 Membership inference attack

MIA [3, 11, 20] are a class of privacy attacks in which an adversary attempts to determine whether a specific data sample was used in training a machine learning model. These attacks exploit differences in the model's behavior when processing training and non-training data, often by leveraging confidence scores or loss values.

Shokri et al. [20] introduced the first large-scale membership inference attack, which utilizes shadow models to approximate the target model's behavior. Numerous studies have shown that overfitting, small training set sizes increase susceptibility to MIA [20, 22]. Specifically, MIA has been shown to be effective even in federated settings, where raw data is not shared but model updates are still vulnerable to inference attacks [11]. Moreover, the clustering strategy in IFCA exacerbates MIA risks, as splitting data into smaller clusters makes minority groups more vulnerable to membership inference.

To counter MIA, various defense mechanisms have been proposed, including differential privacy (DP) [1], adversarial regularization [16], and knowledge distillation [19]. However, these methods often come at the cost of reduced model accuracy and are not specifically designed to address the unique challenges of PFL. Consequently, there is a need for privacy-preserving PFL frameworks that balance accuracy and privacy.

3 Preliminaries

This section provides the foundational background necessary to understand the proposed IFCA-MIR algorithm, including an overview of PFL using IFCA, fairness metrics in machine learning, and the MIA accuracy metric used to evaluate privacy risks.

3.1 Personalized FL with IFCA

In IFCA, the learning problem is framed as a stochastic optimization problem. The objective is to find a set of optimal model parameters $\theta_j^* \in \mathbb{R}^d$ for each cluster $j \in [s]$ such that

$$F(\theta_j) = \mathbb{E}_{z\sim\mathcal{D}_j}\left[f(\theta_j; z)\right], \tag{1}$$

where $f(\theta_j; z)$ is is the local loss function evaluated on data point z under model parameters θ_j and Π_j denotes the underlying data distributions for cluster j which are not directly accessible. Instead, they are accessed through a client datasets $Z_c = \{z|z \sim \Pi_j, z \in \mathbb{D}\}$ where c denotes a client and $\mathbb{D}$ is the domain of data points. The goal is to estimate the membership of each client c to one of the clusters and minimize the empirical loss:

$$\tilde{F}(\theta_j) = \frac{1}{|S_j|}\sum_{c\in S_j}\tilde{F}_c(\theta_j; Z_c), \tag{2}$$

$$\tilde{F}_c(\theta_j; Z_c) = \frac{1}{|Z_c|} \sum_{z_i \in Z_c} f(\theta_j; z_i), \tag{3}$$

where S_j denotes the set of clients assigned to cluster j. The optimization objective is to find the optimal model parameters for each cluster:

$$\tilde{\theta} j^* = \arg\min_{\theta_j} \tilde{F}(\theta_j). \tag{4}$$

3.2 Fairness

Fairness has emerged as a critical concern in machine learning, especially in federated learning scenarios where data is distributed across diverse client groups [5, 15, 17, 23]. In this study, we evaluate fairness using three commonly accepted metrics: Demographic Parity [4], Equal Opportunity [10], and Equalized Odds [10]. We use the notation $\hat{Y} = 1$, $\hat{Y} = 0$ to indicate positive and negative predictions, respectively, and $S = 1$, $S = 0$ to denote the privileged and unprivileged groups.

Definition 1. Demographic Parity is requires that the model's predictions be independent of sensitive attributes. Formally, it is defined as:

$$\mathbb{P}\left[\hat{Y} = 1 | S = 1\right] = \mathbb{P}\left[\hat{Y} = 1 | S = 0\right] \tag{5}$$

Definition 2. Equal Opportunity ensures that true positive rates are equal across groups, promoting fairness without sacrificing accuracy:

$$\mathbb{P}\left[\hat{Y} = 1 | Y = 1, S = 1\right] = \mathbb{P}\left[\hat{Y} = 1 | Y = 1, S = 0\right] \tag{6}$$

Definition 3. Equalized Odds require equal true positive rates and false positive rates across groups:

$$\mathbb{P}\left[\hat{Y} = 1 | Y = y, S = 1\right] = \mathbb{P}\left[\hat{Y} = 1 | Y = y, S = 0\right] \tag{7}$$

This ensures balanced accuracy across both positive and negative outcomes.

In practice, achieving perfect equality in these metrics is challenging. Therefore, the goal is to minimize the absolute difference between the privileged and unprivileged groups, ensuring equitable model performance.

3.3 Merbership inference attack accuracy

We measure MIA accuracy using the following metric:

$$\text{MIA Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} \tag{8}$$

where the True Positive Rate (TPR) and True Negative Rate (TNR) are defined as follows:

$$\text{TPR} = \frac{|\{x \in D_{\text{train}} \mid \hat{y}(x) = 1\}|}{|D_{\text{train}}|} \tag{9}$$

$$\text{TNR} = \frac{|\{x \in D_{\text{non-train}} \mid \hat{y}(x) = 0\}|}{|D_{\text{non-train}}|} \tag{10}$$

Here, D_{train} represents the set of training samples, and $D_{\text{non-train}}$ represents the set of non-training samples. The function $\hat{y}(x)$ denotes the MIA classifier's prediction, where $\hat{y}(x) = 1$ indicates that the sample is inferred as part of the training set, and $\hat{y}(x) = 0$ indicates it is inferred as non-training data.

4 Iterative Federated Clustering Algorithm with Membership Inference Robustness

4.1 Motivation

As mentioned from the above, while IFCA offers advantages in terms of both accuracy and fairness, it becomes more vulnerable to MIA. This vulnerability arises from IFCA's clustering strategy, where the entire training dataset is divided into smaller groups, and models are trained separately for each group. The accuracy of MIA is inversely proportional to the size of the training dataset, with minority groups being particularly more weak to such attacks.

Figure 1 shows the results of performing MIA using Shokri et al.'s method [20] after applying the IFCA technique to the MNIST dataset. Following the experimental setup in [6, 7], we configured the MNIST dataset into majority and minority groups with different distributions. When the proportion of minority participants was set to 10% out of the 200 total participants, the MIA accuracy for the minority group reached 81%, whereas the MIA accuracy for the majority group was 51.5%. When the minority group constituted 30% of the participants, the MIA accuracy dropped to 63%, and when both groups had an equal 50% split, the MIA accuracy was 52%. This demonstrates that MIA accuracy does not significantly increase beyond a certain dataset size, but for datasets below that threshold, the smaller the group, the more vulnerable it becomes to MIA.

To address this disadvantage, we propose an improved version of IFCA, called IFCA-MIR, which incorporates MIA accuracy into the cost function alongside empirical loss. This new algorithm allows personalized models to be trained for each group, minimizing the risk of MIA while maintaining model performance.

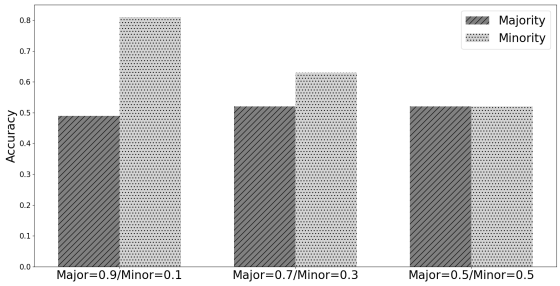


Figure 1: MIA accuracy according to training dataset size

4.2 Notations and problem formulation

Notations. Let $C = \{C_1, \ldots, C_n\}$ be a set of n clients that wish to collaboratively train a model in a federated environment in the presence of an aggregating entity (e.g., a server). Let the clients in C be holding data that are partitioned into s distributions $\Pi_1, \ldots \Pi_s$ for some $s \leq n$. For every $i \in [n]$ and $j \in [s]$, let Z_i be the dataset held by C_i where the datapoints are sampled from the space $\mathcal{Z}$ and let $S^*_{[j]}$ be the set of clients whose data are sampled from Π_j, *i.e.*, $S^*_{[j]} = \{i : z \sim \Pi_j \, \forall \, z \in Z_i\}$. Finally, in any round t, let $\theta^{(t)}_{c[j]}$

denotes the local model of $c \in C$ optimized for cluster j and let $S_{[j]}^{(t)}$ be the set of all clients in round t who reported their optimal cluster ID as j. Hence, let $\theta_{[j]}^{(t+1)} = \frac{1}{\left|S^{(t)[j]}\right|} \sum_{c \in S_{[j]}^{(t)}} \theta_{c[j]}^{(t)}$ denote the aggregated optimized model for cluster j in round t.

Problem formulation. Let $f : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}_{\geq 0}$, where d is the dimension of the model space, be the loss function that the clients seek to minimize for each cluster by finding an optimal model.

In practice, each client holding data sampled from one of the s distributions uses a small sample of their datapoints to evaluate and minimize the average loss they incur on the model in each round to update their local model. Thus, for any finite *batch* $\xi \subset \mathcal{Z}$, the average loss on that batch ξ for any model $\theta \in \mathbb{R}^d$ is given by $f(\theta, \xi) = \frac{1}{|\xi|} \sum_{z \in \xi} f(\theta, z)$. Moreover, let $\ell(\theta, Z)$ denote the accuracy of MIA on θ for any training dataset Z. Therefore, the primary training objective of IFCA-MIR can be formally expressed as:

$$\forall\, j \in [s]:$$
$$\theta_{[j]}^{*} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i \in S_{[j]}^{*}} \frac{1}{|S_{[j]}^{*}|} (\alpha_i f(\theta, Z_i) + \beta_i \ell(\theta, Z_i)), \qquad (11)$$

where, for every $i \in [n]$, $\alpha_i \in [0, 1]$ and $\beta_i \in [0, 1]$ with $\alpha_i + \beta_i = 1$ are hyperparameters chosen by C_i based on its preference on accuracy or privacy.

Remark 1. A high value of α_i and a low value of β_i would exemplify the case where C_i has a greater priority for its model accuracy and less for being protected from a potential MIA. An extreme case of $\alpha_i = 1$ and $\beta_i = 0$ for every $i \in [n]$ reduces down to IFCA.

4.3 Proposed Method: IFCA-MIR

The core idea of IFCA-MIR is that the cluster selection considers not only the empirical loss but also the MIA accuracy. For instance, supposed a client C_i has a local dataset z_i and must choose between models θ_j and θ_k. Even if θ_j yields a lower empirical loss on z_i, it may also exhibit a higher MIA accuracy, indicating greater vulnerability to membership inference attacks. In such a case, a privacy-sensitive client might prefer θ_k, which has a higher empirical loss but offers better privacy protection.

Although IFCA-MIR cannot directly reduce the risk of MIA itself, it enables clients to balance privacy protection and model performance by considering both privacy preferences and empirical loss during cluster selection. To achieve this, we propose the following enhancements in the IFCA-MIR algorithm:

- Red team role for MIA evaluation: The server acts as a red team, simulating external attackers by performing MIA on each θ. In this process, the server calculates the MIA accuracy to assess the privacy vulnerability of each θ, indicating how susceptible each θ is to MIA.
- Providing privacy evaluation information: When the server distributes the θ to clients, it provides not only the candidate θ but also the MIA accuracy information for each θ. Clients can use this information to assess the privacy risks associated with each θ.
- Client's optimal cluster selection: Clients evaluate the empirical loss of each θ received from the server, as done in the original IFCA. However, in IFCA-MIR, clients also consider the MIA accuracy of each θ based on their privacy sensitivity. This approach allows each client to select the θ that strikes the best balance between privacy and performance for their specific needs.

Algorithm 1: IFCA-MIR

Input: Clients $C = \{C_1, \ldots, C_n\}$, Initial models $\{\theta_{[1]}^{(0)}, \ldots, \theta_{[s]}^{(0)}\}$, Learning rate η, Number of rounds T

Output: Optimized models $\{\theta_{[1]}^{(T)}, \ldots, \theta_{[s]}^{(T)}\}$

1 // Server-side
2 Randomly initialize $\theta_{[1]}^{(0)}, \ldots, \theta_{[s]}^{(0)}$
3 **for** $t = 0, \ldots, T-1$ **do**
4 // Train MIA models and compute MIA accuracy
5 **for** $j = 1, \ldots, s$ **do**
6 Compute MIA risk $\ell(\theta_{[j]}^{(t)})$ using **TrainMIA**$(\theta_{[j]}^{(t)})$
7 Send $\{(\theta_{[1]}^{(t)}, \ell(\theta_{[1]}^{(t)})), \ldots, (\theta_{[s]}^{(t)}, \ell(\theta_{[s]}^{(t)}))\}$ to selected clients
8 // Client-side
9 **foreach** *Client* $C_i \in C$ *in parallel* **do**
10 Sample mini-batch ξ_i from local dataset Z_i
11 // Select best model based on empirical loss and MIA risk
12 $\hat{j}_i = \arg\min_{j \in [s]} \alpha_i f(\theta_{[j]}^{(t)}, \xi_i) + \beta_i \ell(\theta_{[j]}^{(t)})$
13 $\tilde{\theta}_{[\hat{j}_i]}^{(t)} \leftarrow \tilde{\theta}_{[\hat{j}_i]}^{(t)} - \eta \nabla f(\tilde{\theta}_{[\hat{j}_i]}^{(t)}, \xi_i)$
14 Send $(\tilde{\theta}_{[\hat{j}_i]}^{(t)}, \hat{j}_i)$ to server
15 // Server-side Aggregation
16 **for** $j = 1, \ldots, s$ **do**
17 $\theta_{[j]}^{(t+1)} = \frac{1}{\left|S^{(t)[j]}\right|} \sum_{c \in S_{[j]}^{(t)}} \theta_{c[j]}^{(t)}$
18 **return** Optimized models $\{\theta_{[1]}^{(T)}, \ldots, \theta_{[s]}^{(T)}\}$

To implement these improvements, we propose the following IFCA-MIR algorithm.

There are two main differences from the original IFCA in this algorithm. First, in IFCA-MIR, the server performs MIA on each model θ (Algorithm 2) and provides the MIA accuracy information to clients, enabling them to make more informed decisions (line 7 in Algorithm 1). Second, clients select the optimal θ^* based on both the empirical loss and the MIA accuracy of the provided θ (line 12 in Algorithm 1).

4.4 Convergence analysis

In this section, we present the theoretical analysis of convergence guarantees of IFCA-MIR. For this, in addition to the notations introduced in Section 4.2, we define the following additional terms. For any model $\theta \in \mathbb{R}^d$ and $j \in [s]$, let $F_{[j]}(\theta)$ denote the *population loss* of cluster j for model θ and is given by the expected loss of θ over the data points following the distribution Π_j, *i.e.*, $F_{[j]}(\theta) = \mathbb{E}_{z \sim \Pi_j}[f(\theta, z)]$. For any client $C_i \in C$ and model $\theta \in \mathbb{R}^d$, we call

Algorithm 2: TrainMIA

Input: model $\theta^{(t)}_{[j]}$, Shadow dataset D_{shadow}

Output: MIA risk score $\ell(\theta^{(t)}_{[j]})$

Train multiple shadow models $\{f^{\text{shadow}}_{\theta^{(t)}_{[j]}}\}$ on D_{shadow}

// Construct attack dataset

foreach *sample z in* D_{shadow} **do**

  Compute confidence scores $S_z = f^{\text{shadow}}_{\theta^{(t)}_{[j]}}(z)$

  Assign label $y = 1$ if $z \in D^{\text{shadow}}_{\text{train}}$, else $y = 0$

  Store (S_z, y) in attack dataset

Train attack model A on the attack dataset (S_z, y)

Compute MIA accuracy $\ell(\theta^{(t)}_{[j]}) = \frac{\text{TPR+TNR}}{2}$

return $\ell(\theta^{(t)}_{[j]})$

$\hat{f}_i(\theta, z) = \alpha_i f(\theta, z) + \beta_i(\theta)\ell(\theta)$ the *privacy-aware loss* of client C_i for model θ computed on data point z. Correspondingly, for all $j \in [s]$ let $\hat{F}_{[j]}$ denote the *population privacy-aware loss* of cluster j given by the average privacy-aware loss of the clients in $S^*_{[j]}$, *i.e.*,

$$\hat{F}_{[j]}(\theta) = \frac{1}{\left|S^*_{[j]}\right|} \sum_{i:\, C_i \in S^*_{[j]}} \mathbb{E}_{z\sim\Pi_j}\left[\hat{f}_i(\theta, z)\right].$$

Let Δ denote the *minimum difference* between the optimal models of any two clusters, *i.e.*, $\Delta = \min_{j\neq j'} \left\|\theta^*_{[j]} - \theta^*_{[j']}\right\|$. Finally, for every $j = 1, \ldots, s$, let $p_j = \left|S^*_{[j]}\right|/n$ denote the fraction of clients belonging to $S^*_{[j]}$ and, hence, $p = \min_{j=1,\ldots,s} p_j$.

In order to proceed with the convergence analysis of IFCA-MIR, we let each client in any given round use a batch of size $B < \infty$ of their local data points to run their local training and we adhere to the same set of assumptions on the loss functions and the initialization conditions that the formal convergence guarantees of IFCA rely on [9].

Assumption 1 (Strong convexity and smoothness of the population privacy-aware loss function). For each $j \in [s]$, the corresponding population loss function $\hat{F}_{[j]}$ is *λ-strongly convex* and *L-smooth*.

Assumption 2 (Bounded variance of privacy-aware loss function). For any $\theta \in \mathbb{R}^d$, $j \in [s]$, and $i \in [n]$, the variance of $\hat{f}_i(\theta, z)$ is bounded by η^2 where $z \sim \Pi_j$, *i.e.*, $\mathbb{E}_{z\sim\Pi_j}\left[(\hat{f}_i(\theta, z) - \hat{F}_{[j]}(\theta, z))^2\right] \leq \eta^2$.

Assumption 3 (Bounded variance on gradients). For any $\theta \in \mathbb{R}^d$, $j \in [s]$, and $i \in [n]$, the variance of $\nabla\hat{f}_i(\theta, z)$ is bounded by σ^2 where $z \sim \Pi_j$, *i.e.*, $\mathbb{E}_{z\sim\Pi_j}\left[\left\|\nabla\hat{f}_i(\theta, z) - \nabla\hat{F}_{[j]}(\theta, z)\right\|_2^2\right] \leq \sigma^2$.

Assumption 4 (Initialization condition). Without the loss of generality, let $\max_{j\in[s]} \left\|\theta^*_{[j]}\right\| \leq 1$. Then for every cluster $j \in [k]$, we assume IFCA-MIR to satisfy the following initialization conditions:

$$\left\|\theta^{(0)}_{[j]} - \theta^*_{[j]}\right\| \leq (\frac{1}{2} - \alpha)\sqrt{\frac{\lambda}{L}}\Delta,$$

such that $0 \leq \alpha \leq \frac{1}{2}$, $B \geq \frac{s\eta^2}{\alpha^2\lambda^2\Delta^4}$, $p \geq \frac{\log(nB)}{n}$, and $\Delta \geq \tilde{O}\left(\max\{\alpha^{-2/5}B^{-1/5}, \alpha^{-1/3}n^{-1/6}B^{-1/3}\}\right)$.

where $\tilde{O}$ any logarithmic factors and terms that do not depend on n and B.

Theorem 4. If Assumptions 1,2,3, and 4 hold, choosing learning rate $\eta = 1/L$, each cluster $j \in [s]$, and any $\delta \in (0, 1)$, in every round $t > 0$, we have with probability at least $(1 - \delta)$:

$$\left\|\theta^{(t+1)}_{[j]} - \theta^*_{[j]}\right\| \leq \left(1 - \frac{p\lambda}{8L}\right)\left\|\theta^{(t)}_{[j]} - \theta^*_{[j]}\right\| + \epsilon_0$$

where $\epsilon_0 \leq \frac{\sigma}{\delta L\sqrt{pnB}} + \frac{\eta^2}{\delta\alpha^2\lambda^2\Delta^4 B} + \frac{\eta\sigma s^{3/2}}{\delta^{3/2}\alpha\lambda L\Delta^2\sqrt{nB}}$.

Corollary 5. If Assumptions 1,2,3, and 4 hold, choosing learning rate $\eta = 1/L$, for each cluster $j \in [s]$ in every round $t > 0$, and any $\delta \in (0, 1)$ and any $\epsilon > 0$, setting $\hat{T} = \frac{8L}{p\lambda}\log\left(\frac{2\Delta}{\epsilon}\right)$, in any round $t \geq \hat{T}$ of IFCA-MIR, we have with probability at least $(1 - \delta)$:

$$\left\|\theta^{(t)}_{[j]} - \theta^*_{[j]}\right\| \leq \epsilon$$

where $\epsilon \leq \frac{\sigma sL\log(nB)}{p^{5/2}\lambda^2\delta\sqrt{nB}} + \frac{\eta^2L^2s\log(nB)}{p^2\lambda^4\delta\Delta^4 B} + \tilde{O}\left(\frac{1}{B\sqrt{n}}\right)$.

The proofs of Theorem 4 and Corollary 5 follow directly from the reasoning presented in the proofs of Theorem 2 and Corollary 2 in [9].

5 Experiments

In this section, we evaluate the proposed IFCA-MIR method and analyze its effectiveness in mitigating MIA risks while maintaining model accuracy and fairness. The experiments were conducted on the MNIST, FEMNIST, and CIFAR-10 datasets, comparing the performance of IFCA-MIR against the original IFCA algorithm.

The objectives of our evaluation are twofold:

- To assess whether IFCA-MIR can reduce MIA vulnerability without compromising model accuracy.
- To examine whether the fairness of the federated learning model is preserved when incorporating MIA robustness into model selection.

By addressing both the privacy and model accuracy aspects in our evaluation, this experiment provides a comprehensive analysis of the IFCA-MIR algorithm's effectiveness in real-world PFL scenarios.

5.1 Datasets

5.1.1 MNIST. The MNIST dataset is a widely-used benchmark for handwritten digit recognition tasks. It consists of grayscale images representing digits 0 through 9, with each image standardized to a size of 28×28 pixels. The dataset includes 60,000 training images and 10,000 test images.

For our experiments, we distributed 50,000 training images evenly among 200 clients for PFL training, while the remaining 10,000 images were used to train a shadow model for MIA analysis on the server.

We preprocessed MNIST based on prior methodologies from [9] and [6]:

- [9]: Rotated MNIST images at different angles to validate personalization.
- [6]: Segmenting the dataset into privileged and unprivileged groups to estimate the fairness.

To simulate real-world non-IID settings, we introduced image rotation variations per client uniformly rather than applying fixed rotation per group:

- Minority group rotations: (0–20°), (0–25°), and (0–30°).
- Majority group rotations: (20–40°), (25–50°), and (30–60°).

By adjusting the distance between the angle boundaries of the two groups, we aim to more accurately simulate real-world data heterogeneity, where distributions often overlap rather than having clearly defined boundaries.

5.1.2 FEMNIST. The FEMNIST (Federated Extended MNIST) dataset is specifically designed to evaluate performance in federated learning environments and exhibits a non-IID data distribution [2]. Derived from the EMNIST dataset, FEMNIST consists of handwritten images representing 62 classes, including digits 0-9 and both uppercase and lowercase alphabetic characters (A-Z, a-z). The dataset comprises approximately 700,000 training samples and 100,000 test samples, collected from around 3,500 clients.

To control the inherent non-IID characteristics of the FEMNIST dataset, we first aggregated the training data, which was originally divided by individual clients, into a unified dataset. We then evenly distributed 600,000 training samples among 200 clients and used the remaining 100,000 samples to train a shadow model for MIA on the server.

We preprocessed the FEMNIST dataset using the same methodology as that used for the MNIST dataset.

- Minority group rotations: (0–20°), (0–25°), and (0–30°).
- Majority group rotations: (25–45°), (30–55°), and (35–65°).

5.1.3 CIFAR-10. The CIFAR-10 dataset consists of 60,000 color images categorized into 10 classes, with each image sized at 32x32 pixels. It contains 50,000 training images and 10,000 test images, uniformly distributed across all classes. For our study, 50,000 training images were evenly allocated to 100 clients for PFL training and 10,000 images were used to train the shadow model for MIA.

Since CIFAR-10 images are more complex than those in MNIST and FEMNIST, controlling the non-IID characteristics through image rotation was not feasible. Instead, we varied the brightness levels to simulate non-IID data distributions.

- Majority group brightness: (0.5-0.8)
- Minority group brightness: (0.8-1.1), (0.85-1.15), and (0.9-1.2).

This approach introduces non-IID conditions while preserving image integrity.

5.2 Experimental Setup

The goal of our experiment is to verify whether the IFCA-MIR can minimize the number of clients exposed to MIA risks without compromising model accuracy, while also maintaining fairness. To achieve this, we designed experiments focusing on three key aspects: model accuracy, MIA vulnerability, and fairness. Each experiment was repeated five times to ensure reliability. Additionally, the server performs MIA evaluations every five iterations with three shadow models rather than at every iteration. The frequency of MIA evaluations and number of shadow models can be adjusted as needed based on specific requirements.

5.2.1 Model Accuracy. To evaluate model accuracy, we used the MNIST, FEMNIST, and CIFAR-10 datasets, comparing the performance of IFCA-MIR against the original IFCA algorithm.

For MNIST and FEMNIST, the total number of clients was set to 200, with each client receiving an equal number of data points. For CIFAR-10, the number of clients was set to 100.

The objective was to compare model accuracy between the majority and minority groups. We assumed the presence of two models for PFL, aligning with the two distinct distributions in each dataset. we compared the accuracy of the original IFCA algorithm with that of IFCA-MIR, examining whether IFCA-MIR could maintain model accuracy without degradation.

5.2.2 MIA vulnerability. To evaluate MIA vulnerability, we used the same datasets and experimental settings as those employed in the model accuracy evaluation. In this evaluation, each client was assigned an MIA vulnerability threshold, and we counted the number of clients whose personalized model's MIA accuracy exceeded these thresholds.

The MIA threshold defines the maximum allowable MIA accuracy, reflecting each client's privacy preference. We uniformly assigned the MIA thresholds between 50% and 80% across all clients. A lower threshold indicates a higher sensitivity to MIA, while a higher threshold suggests a greater tolerance for MIA risks.

For example, if a client sets an MIA threshold of 60% but is assigned a personalized model with an MIA accuracy of 65%, their threshold would be exceeded. Conversely, a client with a threshold of 70% in the same group would remain within their acceptable limit. This approach enabled us to effectively evaluate the extent to which each client's privacy preferences were upheld.

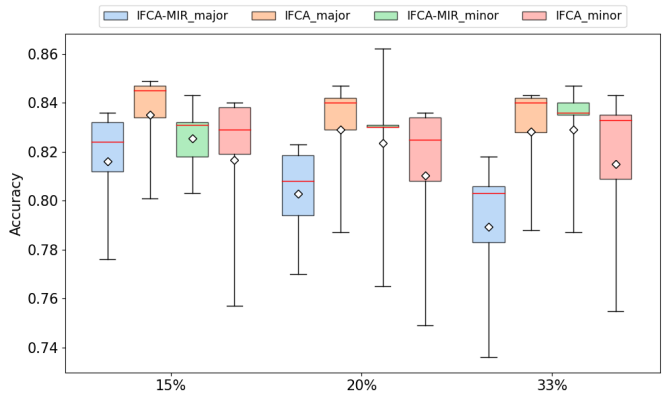
5.2.3 Fairness. To evaluate fairness, we utilized three metrics: demographic parity, equal opportunity, and equalized odds. We compared the results between the majority and minority groups to assess whether fairness was preserved in the proposed method. This analysis allowed us to determine whether the IFCA-MIR algorithm maintains fairness while simultaneously mitigating MIA vulnerability compared to the original IFCA algorithm.

5.3 Model Accuracy

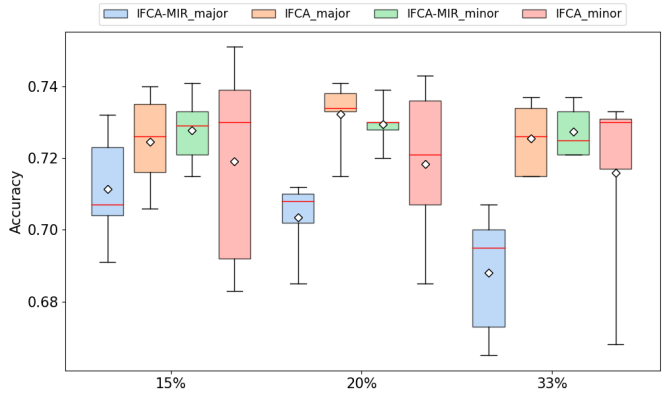
We evaluated the model accuracy of both the original IFCA and the proposed IFCA-MIR by varying image rotation degrees (for MNIST and FEMNIST) and brightness levels (for CIFAR-10), as well as by adjusting the proportion of the minority group relative to the majority group. Specifically, we compared the accuracy of personalized models for both the majority and minority groups under the original IFCA and IFCA-MIR methods.

Figure 2 illustrates the results when the proportion of the minority group was varied. For MNIST, the majority group's image rotation angles were fixed within the range of (25-50) degrees, while the minority group's angles were set within (0-25) degrees. For FEMNIST, the majority group's rotation angles were fixed within (30-55) degrees, while the minority group's angles were within (0-25) degrees. In CIFAR-10, the majority group's brightness was set within

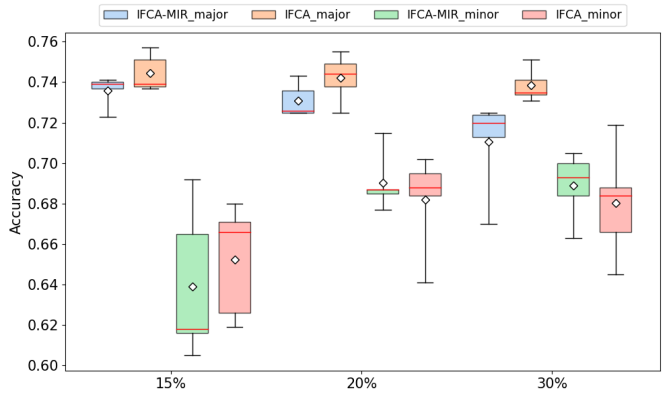
the range of (0.85-1.15), whereas the minority group's brightness was fixed within (0.5-0.8).



(a) MNIST



(b) FEMNIST



(c) CIFAR-10

Figure 2: Model accuracy results for each dataset with varying minority dataset sizes. The white diamond represents the mean, while the red line within the box plot indicates the median

Table 1: Comparison of Average Model Accuracy Across Different Datasets

Accuracy	MNIST	FEMNIST	CIFAR-10
IFCA [9]	0.816	0.727	0.725
IFCA-MIR	0.797	0.704	0.719

The results indicate that for the majority group, IFCA-MIR yielded slightly lower accuracy compared to the original IFCA. However, this difference was not substantial enough to indicate a significant performance drop. In contrast, the minority group's accuracy was either comparable to or improved over the original IFCA. This improvement can be attributed to IFCA-MIR 's incorporation of both MIA vulnerability and loss, which led some clients originally classified as minority to migrate to the majority group to reduce their MIA risks.

As a result, the remaining minority group consisted of clients who were willing to accept a higher MIA risk in exchange for better model accuracy. Consequently, their accuracy either remained equivalent to or exceeded that of the original IFCA. This trend was consistently observed across all datasets, demonstrating that IFCA-MIR maintains stable performance across different data distributions.

Figure 3 presents the results when the degree of image deformation was modified while keeping the minority client ratio fixed. Specifically, the minority ratio was fixed at 15% for MNIST and FEMNIST, and at 20% for CIFAR-10. The experiment was conducted by varying image rotation and brightness levels. The overall trend was consistent with the previous results. However, for CIFAR-10, the accuracy of the minority model under IFCA-MIR decreased as the boundary between the majority and minority group distributions became less distinct. This is because, as the distributional differences between the two groups narrowed, more minority clients migrated to the majority group, reducing the amount of data available for training the minority model. Nevertheless, even in these cases, the decline in accuracy was minimal and remained comparable to that of the original IFCA. Additionally, as the range of image deformation increases, the training complexity rises, resulting in an overall decrease in accuracy.

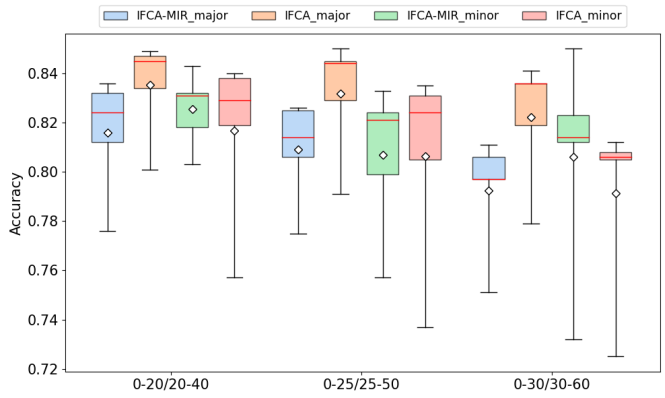
Table 1 presents the overall accuracy averages, which combine the accuracies of both the Major and Minority models. As shown in the tables, despite the differences between the Major and Minority models, the overall model accuracies of the original IFCA and IFCA-MIR remain comparable.

5.4 MIA vulnerability

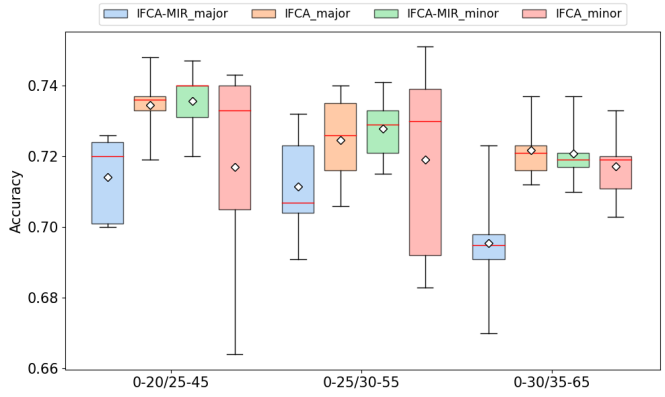
We conducted a comparative evaluation of MIA vulnerability between the original IFCA and IFCA-MIR methods, using the same experimental setup as the model accuracy experiments.

Table 2 presents the MIA accuracy of the Majority and Minority models for each dataset, comparing the original IFCA with IFCA-MIR. If a client's MIA threshold is lower than this accuracy, it is considered a violation of MIA.

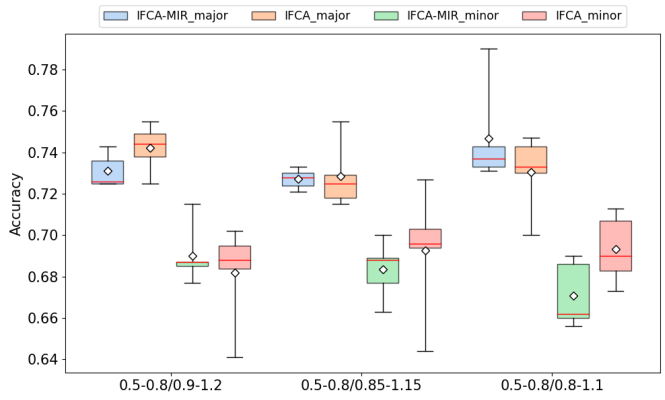
Figures 4 and 5 illustrate the results of this comparison. The experimental findings indicate that IFCA-MIR significantly reduces



(a) MNIST



(b) FEMNIST



(c) CIFAR-10

Figure 3: Model accuracy results for each dataset with varying image deformation ranges

MIA vulnerability compared to the original IFCA. This improvement can be attributed to the fact that the proposed method enables privacy-sensitive clients to select a safer group rather than solely optimizing for loss. Our experiments demonstrate that even without additional weight optimization, allowing clients to make group selections based on their privacy preferences leads to improved outcomes. This highlights the importance of integrating privacy considerations alongside traditional performance metrics, such as loss, in personalized model training. The results confirm that IFCA-MIR effectively reduces the number of clients exposed to MIA risks.

By considering both loss and privacy concerns, the proposed method provides a flexible and adaptive framework that allows clients to balance performance and privacy based on their individual preferences.

Table 2: Comparison of Average MIA accuracy Across Different Datasets

		MNIST	FEMNIST	CIFAR-10
IFCA [9]	Major	0.521	0.537	0.604
	Minor	0.761	0.722	0.785
IFCA-MIR	Major	0.530	0.525	0.595
	Minor	0.791	0.733	0.792

5.5 Fairness

We conducted experiments to evaluate the fairness of the proposed method using the MNIST and FEMNIST datasets. As discussed earlier, PFL has been shown to enhance fairness compared to centralized learning methods [6]. In this experiment, we aimed to assess the impact of the proposed IFCA-MIR algorithm on fairness in comparison to the original IFCA approach.

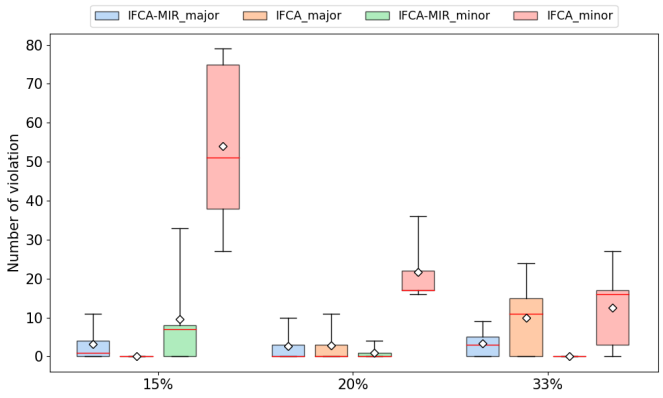
The experimental results indicate that IFCA-MIR does not introduce differences in fairness compared to the original IFCA method. Notably, on the MNIST dataset, IFCA-MIR demonstrated slightly improved fairness over the original IFCA. The absolute fairness values for both methods remained close to zero, suggesting that both IFCA and IFCA-MIR contribute to fairer outcomes in FL environments. Overall, the findings demonstrate that IFCA-MIR preserves the fairness advantages of PFL without introducing substantial trade-offs compared to the original version of IFCA. In the interest of space, the experimental results for fairness are presented in Appendix A.

5.6 Convergence

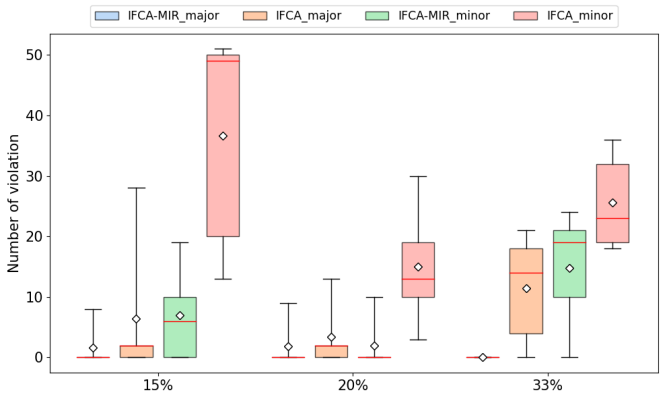
We derive the theoretical convergence guarantee of IFCA-MIR in Section 4.4 and validate it experimentally. As shown in Figure 6, the models for both the majority and minority groups trained using IFCA-MIR successfully converge across all datasets.

5.7 Discussion

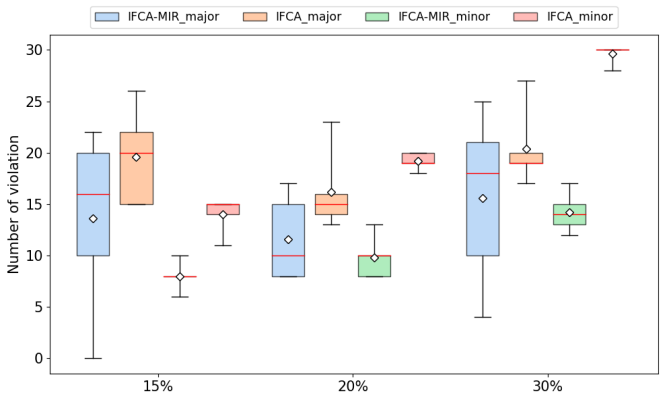
We show that IFCA-MIR successfully reduces MIA accuracy while maintaining overall model performance and fairness. However, during this balancing process, we observed a shift of minority group clients to the majority group. This phenomenon arises because the majority group generally offers better privacy protection compared to the minority group. When the difference in accuracy between the majority and minority groups is small, privacy-sensitive clients



(a) MNIST

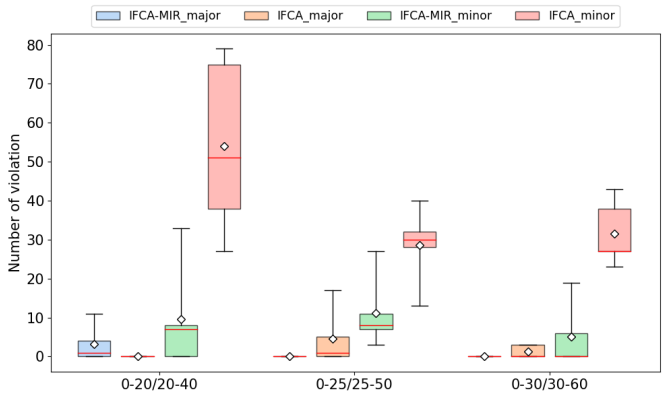


(b) FEMNIST

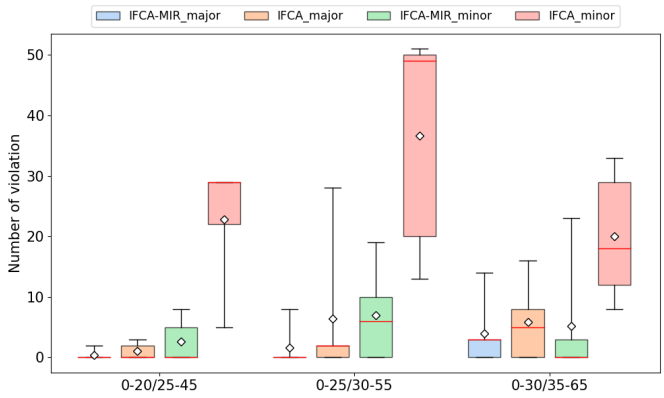


(c) CIFAR-10

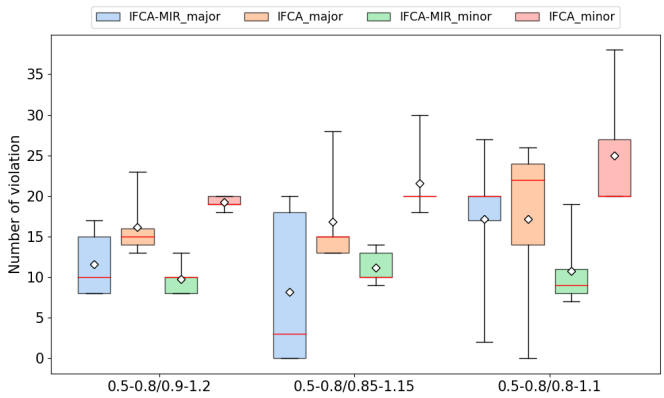
Figure 4: Number of MIA violation for each dataset with varying minority dataset sizes



(a) MNIST



(b) FEMNIST



(c) CIFAR-10

Figure 5: Number of MIA violation for each dataset with varying image deformation ranges

accept a slight accuracy loss to migrate to the majority group, as it provides lower MIA risk.

While this trade-off is a fundamental aspect of our method, it also introduces a potential over-concentration in the majority group, which can significantly degrade the performance of the minority group. As shown in Figure 2c, such shifts may reduce the accuracy of the minority group, and it might even make it reach an unacceptable level. From a global optimization perspective, the server may need additional protocols to prevent excessive client migration to the majority group. This remains a promising direction for future research.

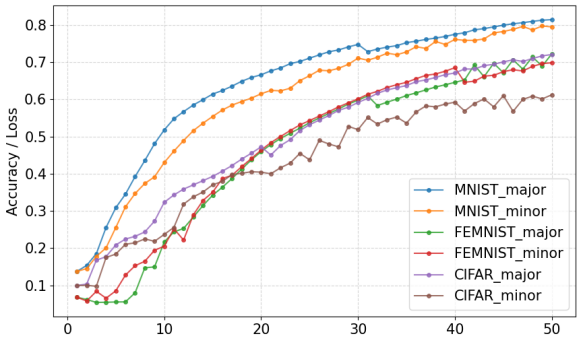


Figure 6: Convergence of majority and minority models for each dataset

Another key consideration in our approach is the reliance on the server to act as a "red team" for MIA evaluation. Our method assumes that the server performs MIA, evaluates its risk, and provides this information to clients, enabling privacy-aware model selection. However, this approach is only valid under the assumption that the server is fully trusted. In settings where the server cannot be trusted, clients may instead perform MIA evaluation locally to ensure privacy-preserving model selection.

6 Conclusion

In this paper, we proposed IFCA-MIR, an improved clustering-based PFL algorithm that mitigates MIA vulnerability while maintaining model accuracy and fairness. Unlike the original IFCA, which selects models based solely on empirical loss, IFCA-MIR incorporates MIA risk assessment into the model selection process. This enables clients to balance privacy protection and model performance based on their individual sensitivity to MIA.

Through extensive experiments on the MNIST, FEMNIST, and CIFAR-10 datasets, we demonstrated that IFCA-MIR effectively reduces the number of clients exposed to MIA risks compared to the original IFCA. Our findings show that while the majority group accuracy remains comparable between both methods, the minority group accuracy is either maintained or improved, as privacy-sensitive clients migrate to safer clusters. Moreover, our fairness evaluation confirmed that IFCA-MIR preserves the fairness benefits of personalized federated learning without introducing significant trade-offs. Overall, IFCA-MIR provides up to 5.6× better performance in mitigating the risk of MIA in FL, while maintaining model accuracy and fairness.

Developing an incentive-based protocol to prevent excessive migration of minority-group clients to majority groups remains a promising direction for future work. To achieve this, the server should act as a coordinator that seeks a Pareto optimality between overall model accuracy and MIA vulnerability. Another interesting avenue to explore in the future is the application of DP in IFCA-MIR. While DP is a well-established technique to defend against MIA, our study focused on reducing MIA vulnerability without relying on adding noise to achieve DP. As introducing DP may lead to a trade-off between privacy protection and model accuracy, exploring optimal DP strategies that minimize accuracy degradation while ensuring robustness against MIA would be valuable for extending the applicability of IFCA-MIR and, more generally, for future research in this domain.

Acknowledgments

This work has been funded by the European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2018. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097* (2018).

[3] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *International conference on machine learning*. PMLR, 1964–1974.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[5] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. Fairfed: Enabling group fairness in federated learning, In 1st NeurIPS Workshop on New Frontiers in Federated Learning. *arXiv preprint arXiv:2110.00857*. https://arxiv.org/abs/1611.04482

[6] Filippo Galli, Kangsoo Jung, Sayan Biswas, Catuscia Palamidessi, and Tommaso Cucinotta. 2023. Advancing personalized federated learning: Group privacy, fairness, and beyond. *SN Computer Science* 4, 6 (2023), 831.

[7] Filippo Galli, Kangsoo Jung, Sayan Biswas, Catuscia Palamidessi, and Tommaso Cucinotta. 2023. Group privacy for Personalized Federated Learning. In *9th International Conference on Information Systems Security and Privacy (ICISSP)*. 252–263.

[8] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems* 33 (2020), 16937–16947.

[9] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. 2020. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems* 33 (2020), 19586–19597.

[10] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[11] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[12] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. 2021. Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1102–1107.

[13] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. 2020. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619* (2020).

[14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[15] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*. PMLR, 107–118.

[16] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

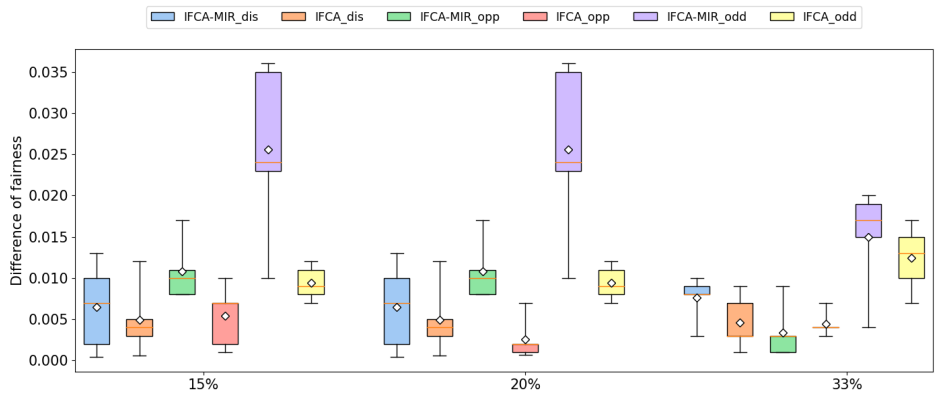
[17] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[18] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. 2020. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems* 32, 8 (2020), 3710–3722.

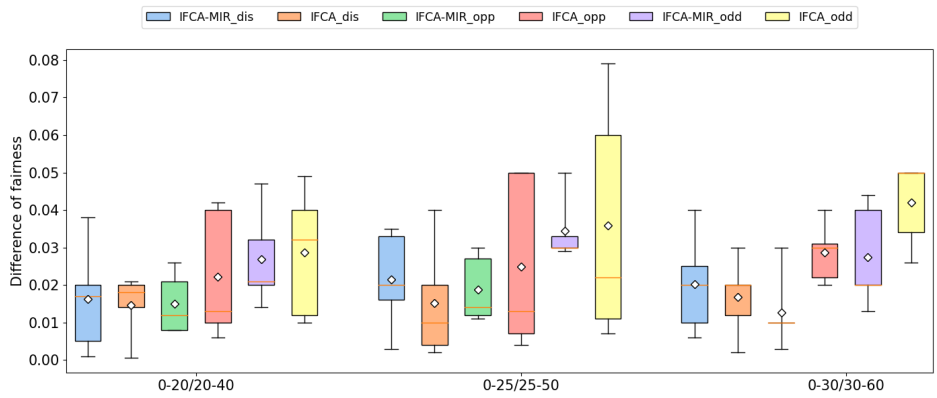
[19] Virat Shejwalkar and Amir Houmansadr. 2021. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 9549–9557.

[20] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
[21] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems* 34, 12 (2022), 9587–9603.
[22] Marlon Tobaben, Hibiki Ito, Joonas Jälkö, Gauri Pradhan, Yuan He, and Antti Honkela. 2024. Impact of dataset properties on membership inference vulnerability of deep transfer learning. *arXiv preprint arXiv:2402.06674* (2024).
[23] Michael Wick, Jean-Baptiste Tristan, et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems* 32 (2019).
[24] Gongxi Zhu, Donghao Li, Hanlin Gu, Yuan Yao, Lixin Fan, and Yuxing Han. 2025. FedMIA: An Effective Membership Inference Attack Exploiting" All for One" Principle in Federated Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20643–20653.
[25] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. *Deep leakage from gradients*. Curran Associates Inc., Red Hook, NY, USA.

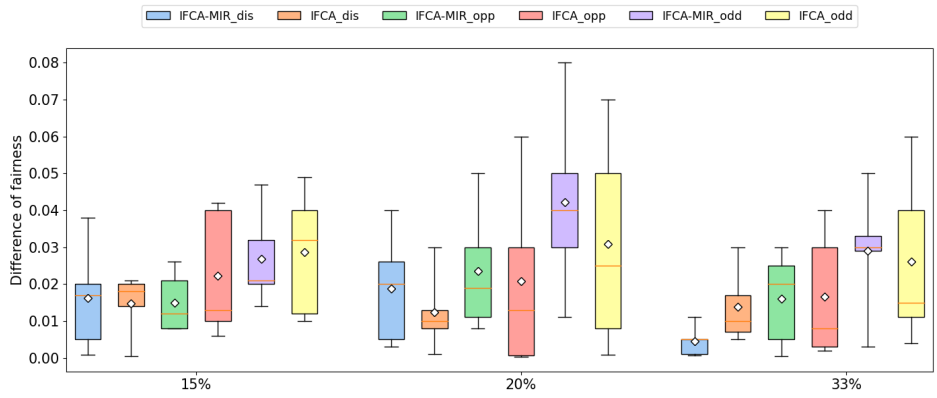
A Experimental results of fairness



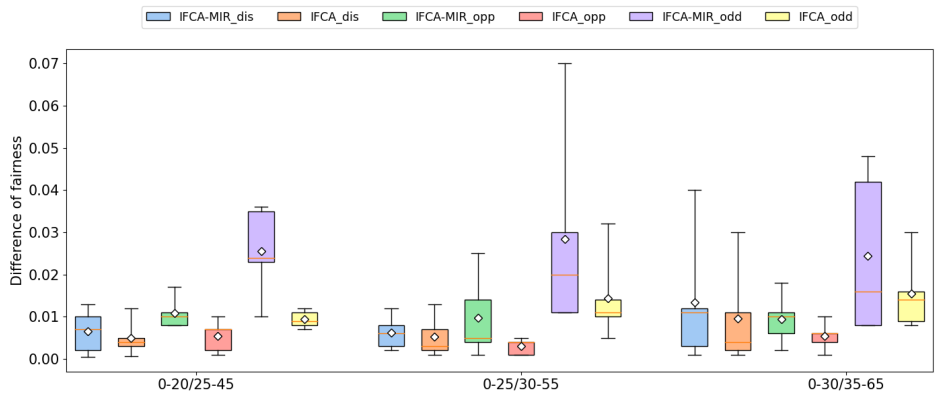
(a) Varying minority dataset sizes for MNIST



(b) Varying image deformation ranges for MNIST



(c) Varying minority dataset sizes for FEMNIST



(d) Varying image deformation ranges for FEMNIST

Figure 7: Fairness comparison for MNIST and FEMNIST dataset with varying minority dataset sizes and image deformation ranges